

Feature Binding with Category-Dependant MixUp for Semantic Segmentation and Adversarial Robustness

Md Amirul Islam^{1,2}

www.cs.ryerson.ca/~amirul

Matthew Kowal¹

mkowal2.github.io

Konstantinos G. Derpanis^{1,2,3}

www.cs.ryerson.ca/~kosta

Neil D. B. Bruce^{1,2,4}

www.cs.ryerson.ca/~bruce

¹ Ryerson University
Toronto, Canada

² Vector Institute
Toronto, Canada

³ Samsung AI Centre
Toronto, Canada

⁴ University of Guelph
Guelph, Canada

Abstract

In this paper, we present a strategy for training convolutional neural networks to effectively resolve interference arising from competing hypotheses relating to inter-categorical information throughout the network. The premise is based on the notion of feature binding, which is defined as the process by which activation’s spread across space and layers in the network are successfully integrated to arrive at a correct inference decision. In our work, this is accomplished for the task of dense image labelling by blending images based on their class labels, and then training a *feature binding* network, which simultaneously segments and separates the blended images. Subsequent feature denoising to suppress noisy activations reveals additional desirable properties and high degrees of successful predictions. Through this process, we reveal a general mechanism, distinct from any prior methods, for boosting the performance of the base segmentation network while simultaneously increasing robustness to adversarial attacks.

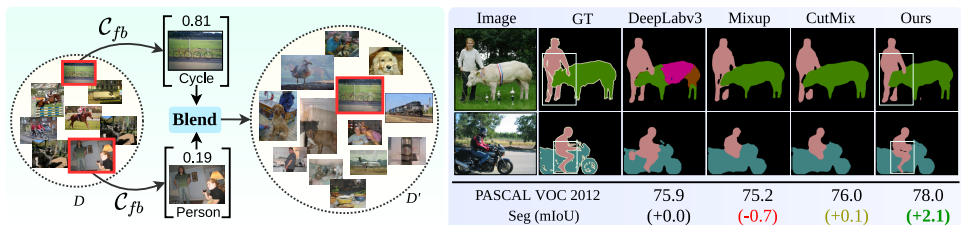


Figure 1: Left: Overview of our category-specific (C_{fb}) image blending to create a new source dataset (D'). A segmentation network is trained with D' to simultaneously separate and segment both source images. Right: Results of our *feature binding* method, Mixup [69], and CutMix [88] on the PASCAL VOC 2012 [11] segmentation task. Note that, our method significantly improves the performance.

1 Introduction

The advent of Deep Neural Networks (DNNs) has seen overwhelming improvement in dense image labeling tasks [2, 6, 13, 15, 18, 19, 20, 21, 22, 23, 24, 25, 27, 30, 40], however, for some common benchmarks [10] the rate of improvement has slowed down. While one might assume that barriers to further improvement require changes at the architectural level, it has also been borne out that pre-training across a variety of datasets [26, 33] can improve performance exceeding improvements seen from changing the model architecture. However, there are challenging scenarios for which DNNs have difficulty on regardless of pre-training or architectural changes, such as highly occluded scenes, or objects appearing out of their normal context [35]. It is not clear though, for dense image labeling tasks, how to resolve these specific scenarios for more robust prediction quality on a per-pixel level.

A question that naturally follows from this line of reasoning is: How can the number of locally challenging cases be increased, or the problem made more difficult in general? In this paper, we address this problem using a principled approach to improve performance and that also implies a more general form of robustness. As inspiration, we look to a paradigm discussed often in the realm of human vision: the binding problem [34, 37]. The crux of this problem is that given a complex decomposition of an image into features that represent different concepts, or different parts of the image, how does one proceed to successfully relate activations corresponding to common sources in the input image to label a *whole* from its parts, or separate objects. Motivated by the binding problem, a successful solution in the computer vision domain should rely on both determining correspondences in activations among features that represent disparate concepts, and also to associate activations tied to related features that are subject to spatial separation in the image. To address similar issues for the image classification task, recent studies [36, 38, 39] have considered mixing two image examples with constraints on the distribution of features. However, these methods suffer from biases in the dataset used, as they have no strategy when deciding on which images to mix which is crucial for the dense labeling problem. Additionally, these strategies do not adequately separate information from different sources in the image as they only require the network to make a single (classification) prediction during training.

In our work, the means of solving the feature binding problem takes a direct form, which involves training networks on a specially designed dataset of mixed images to simultaneously address problems of dense image labeling [8, 27, 30], and blind source separation [12, 16]. Humans show a surprising level of capability in interpreting a superposition (e.g., average) of two images, both interpreting the contents of each scene and determining the membership of local patterns within a given scene. The underlying premise of this work involves producing networks capable of simultaneously performing dense image labeling for pairs of images while also separating labels according to the source images. If one selects pairs on the basis of a weighted average (see Fig. 1 (left)), this allows treatment of the corresponding dense image labeling problem in the absence of source separation by extension. This process supports several objectives: (i) it significantly increases the number of occurrences that are locally ambiguous that need to be resolved to produce a correct categorical assignment, (ii) it forces broader spatial context to be considered in making categorical assignments, and (iii) it stands to create more powerful networks for standard dense labeling tasks and dealing with adversarial perturbations by forcing explicit requirements on how the network uses the input. The end goal of our procedure is to improve overall performance as well as increase the prediction quality on complex images (see Fig. 1 (right)), heavily occluded scenes, and also invoke robustness to challenging adversarial inputs. Our main contributions are as follows:

- To the best of our knowledge we present the first work which applies image blending to the dense labeling task. To this end, we propose a novel training pipeline which simultaneously solves the problems of dense labeling and blind source separation.
- We further introduce a new *categorical clustering* strategy which exploits semantic knowledge of the dataset to mix input images based on their class distributions.
- We show, through extensive quantitative and qualitative experiments, that our pipeline outperforms recent image blending methods [58, 59] on the PASCAL VOC 2012 dataset [10], while simultaneously improving robustness to adversarial attacks.

2 Related Work

More closely related to the feature binding concept, contributions [0, 8, 11, 12, 17, 36, 58, 59] on data augmentation based techniques share a similar idea of mixing two randomly selected samples to create new training data for the image classification or localization task. BC learning [36] showed that randomly mixing training samples can lead to better separation between categories based on the feature distribution. Mixup [59] shares a similar idea of training a network by mixing the data that regularizes the network and increases the robustness against adversarial examples, whereas CutMix [58] proposed to overlay a cropped area of an input image to another. Our proposed *feature binding* approach differs from the above existing works in that: (i) the network performs simultaneous dense prediction and blind source separation to achieve superior dense labeling and adversarial robustness whereas other techniques are focused mainly on image classification or object localization, (ii) previous methods either mix labels as the ground truth or use the label from only one sample, while we use both ground truth labels independently, and (iii) samples are chosen randomly for Mixup [59] and CutMix [58] while we use an intuitive strategy (categorical clustering, Sec. 3.1).

3 Proposed Method

In the broader context of investigating approaches motivated by the feature binding problem, we propose a novel framework capable of solving the dense labeling problem. Our proposed framework consists of three key steps: (i) we first apply a technique on the training dataset that generates a new set of source images (Sec. 3.1), (ii) we train a convolutional neural network (CNN) using the generated data that produces dense predictions (Sec. 3.2), and (iii) we denoise the learned features from the feature binding process by fine-tuning on the standard data (Sec. 3.3).

3.1 Category-Dependent Image Blending

Recent works [8, 17, 36, 58, 59] simply mix two randomly selected samples to create new training data for classification or localization task. Exploring a similar direction, we are interested in solving dense prediction in a way that provides separation based on mixed source images. We augment the PASCAL VOC 2012 [10] training dataset via a novel data processing stage to generate a new training set in a form that accounts for source separation and dense prediction. The traditional way [17, 36, 59] of combining two images is by weighted average which implies that the contents of both scenes appear with varying contrast. Randomly combining two source images to achieve the desired objective is a more significant

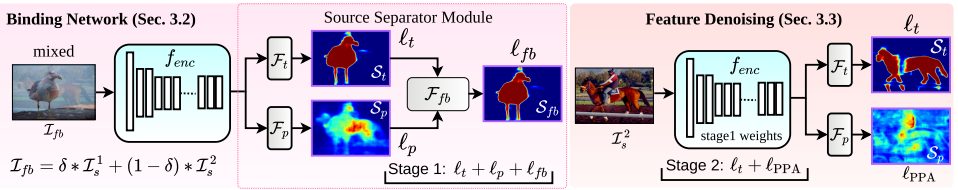


Figure 2: An illustration of *feature binding* process. At the data end, categorical collisions are created with a dominant (\mathcal{I}_s^1) and phantom (\mathcal{I}_s^2) image. **Stage 1:** The network is trained on mixed data (\mathcal{I}_{fb}) to perform simultaneous dense labeling and source separation. We use the labels of both source images as the targets for two separate output channels. **Stage 2:** Fine tuning on standard data to further promote desirable properties along the two dimensions of base performance and robustness to perturbations. In this stage, the phantom activation of the second channel is suppressed. Confidence maps are plotted with the ‘Jet’ colormap, where red and blue indicates higher and lower confidence, respectively.

challenge than one might expect in the context of dense prediction. One challenge is the categorical bias of the dataset (e.g., mostly the *person* images will be combined with all other categories since *person* is the most common category in PASCAL VOC 2012) across the newly generated training set. Previous methods [17, 37, 39], randomly select images to combine, results in a new data distribution which inherit similar biases as the original dataset. To overcome these limitations, we propose a technique denoted as *categorical clustering*, \mathcal{C}_{fb} , which combines images based on a uniform distribution across categories. Thorough experimentation with our proposed mixing strategy show improvements in the network’s ability to separate competing categorical features and can generalize these improvements to various challenging scenarios, such as segmenting out-of-context objects or highly occluded scenes. **Categorical Clustering:** We first generate 20 different clusters of images where each cluster contains images of a certain category from VOC 2012. For each training sample in a cluster, we linearly combine it with a random sample from each of the 19 other clusters. For example, given a training sample \mathcal{I}_s^1 from the *person* cluster we randomly choose a sample \mathcal{I}_s^2 from another categorical cluster and combine them to obtain a new sample, \mathcal{I}_{fb} :

$$\mathcal{I}_{fb} = \delta * \mathcal{I}_s^1 + (1 - \delta) * \mathcal{I}_s^2, \quad (1)$$

where δ denotes the randomly chosen weight that is applied to each image. We assign the weight such that the source image (\mathcal{I}_s^1) has more weight compared to the random one (\mathcal{I}_s^2). In our experiments, we sample δ uniformly from a range of $[0.7 - 1]$ for each image pair. We also change the range of δ values and report results in Table 5 (a). Note that, for one sample in *person* cluster we generate 19 new samples. We continue to generate feature binding samples for the other remaining images in the person cluster and perform the same operation for images in other clusters. While there may exist alternatives [9, 38, 39] for combining pairs of images to generate a training set suitable for source separation training, our intuitive method is simple to implement and achieves strong performance on a variety of metrics (see Section 4). Exploring further methods to combine and augment the training set is an interesting and nuanced problem to be studied further in dense image labelling.

3.2 Feature Binding Network

In this section, we present a fully convolutional feature binding network in the context of dense prediction. Fig. 2 illustrates the overall pipeline of our proposed method. Fig. 2 (left) reveals two key components of the binding network including a *fully convolutional*

network encoder and *source separator module* (SSM). Given a mixed image $\mathcal{I}_{fb} \in \mathbb{R}^{h \times w \times c}$, we adopt a DeepLabv3 [9] (f_{enc}) to produce a sequence of bottom-up feature maps. The SSM consists of two separate branches: (i) *dominant* $\mathcal{F}_t(\cdot)$, and (ii) *phantom*, $\mathcal{F}_p(\cdot)$. Each branch takes the spatial feature map, \hat{f}_b^i , produced at the last block, $r \in \text{set5c}$, of f_{enc} as input and produces a dense prediction for the dominant, \mathcal{S}_t , and the phantom, \mathcal{S}_p , image. Next, we append a *feature binding head* (FBH) to generate a final dense prediction of categories for the dominant image. The FBH, \mathcal{F}_{fb} , simply concatenates the outputs of source and phantom branches followed by two 1×1 convolution layers with non-linearities (ReLU) to obtain the final dense prediction map, \mathcal{S}_{fb} . The intuition behind the FBH is that the phantom branch may produce activations that are correlated with the dominant image, and thus the FBH allows the network to further correct any incorrectly separated features with an additional signal to learn from. Given a mixed image, \mathcal{I}_{fb} , the operations can be expressed as:

$$\hat{f}_b^i = f_{enc}(\mathcal{I}_{fb}), \quad \underbrace{\mathcal{S}_t = \mathcal{F}_t(\hat{f}_b^i)}_{\text{dominant}}, \quad \underbrace{\mathcal{S}_p = \mathcal{F}_p(\hat{f}_b^i)}_{\text{phantom}}, \quad \underbrace{\mathcal{S}_{fb} = \mathcal{F}_{fb}(\mathcal{S}_t, \mathcal{S}_p)}_{\text{binding}}. \quad (2)$$

Training the Feature Binding Network. The feature binding network produces two dominant predictions, \mathcal{S}_{fb} and \mathcal{S}_t , including a phantom prediction, \mathcal{S}_p ; however, we are principally interested in the final dominant prediction, \mathcal{S}_{fb} . In more specific terms, let $\mathcal{I}_{fb} \in \mathbb{R}^{h \times w \times 3}$ be a training image associated with ground-truth maps ($\mathcal{G}_s^1, \mathcal{G}_s^2$) in the feature binding setting. To apply supervision on \mathcal{S}_{fb} , \mathcal{S}_t , and \mathcal{S}_p , we upsample them to the size of \mathcal{G}_s^1 . Then we define three pixel-wise cross-entropy losses, ℓ_{fb} , ℓ_t , and ℓ_p , to measure the difference between $(\mathcal{S}_{fb}, \mathcal{G}_s^1)$, $(\mathcal{S}_t, \mathcal{G}_s^1)$, and $(\mathcal{S}_p, \mathcal{G}_s^2)$, respectively. The objective function can be formalized as:

$$L_{stage1} = \ell_{fb} + \delta * \ell_t + (1 - \delta) * \ell_p, \quad (3)$$

where δ is the weight used in to linearly combine images to generate \mathcal{I}_{fb} . Note that the network is penalized the most on the final and initial dominant predictions, and places less emphasis on the phantom prediction.

3.3 Denoising Feature Binding

While feature binding and source separation are interesting, the ultimate goal is to see improvement and robustness for standard images. For this reason, we mainly care about improving the overall dense prediction. To accomplish this, we further fine-tune our trained binding model on the standard training set which we call the feature denoising stage. In this stage, as we feed a standard image to the network, the phantom predictor branch, \mathcal{F}_{ph} , has no supervisory signal, instead it acts as a regularizer. We propose the following technique to penalize the phantom prediction.

Penalize Phantom Activation: Along with ℓ_t , we propose a loss, ℓ_{PPA} , on the phantom prediction to penalize any activation (and suppress phantom signals and interference). The goal here is to push the output of the phantom branch to zero and getting rid of the phantom. The ℓ_{PPA} loss sums the absolute value of the confidence attached to categories and applies a log operation to balance the numeric scale with ℓ_t :

$$\ell_{PPA} = \log \sum_{\forall i \in h} \sum_{\forall j \in w} \sum_{\forall k \in c} \sigma(\mathcal{S}_p), \quad L_{stage2} = \ell_t + \ell_{PPA}, \quad (4)$$

*	Method	mIoU (%)
Val	DeepLabv3-ResNet101 [9]	75.9
	DeepLabv3 + Mixup [39]	75.2
	DeepLabv3 + CutMix [38]	76.0
	DeepLabv3 + Feature Binding	78.0
Test	DeepLabv3 [9]	79.3
	DeepLabv3 + Feature Binding	82.1

Table 1: (a) PASCAL VOC 2012 val and test set results for the baselines and our approach.

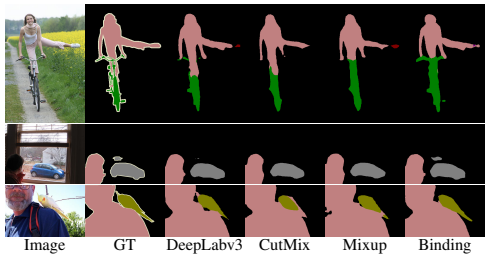


Figure 3: Qualitative results on the PASCAL VOC 2012 validation set.

where $\sigma(\cdot)$ is the ReLU function, which constrains the input to the log to be a positive value. In **Stage 1**, f_{enc} , \mathcal{F}_i , \mathcal{F}_p , and \mathcal{F}_{fb} are trained in an end-to-end manner. Then, in **Stage 2**, f_{enc} , \mathcal{F}_i , and \mathcal{F}_p are fine-tuned from the Stage 1 weights.

4 Experiments

We first present results on the PASCAL VOC 2012 [10] semantic segmentation dataset (Sec. 4.1). Unless otherwise stated, we use the DeepLabv3 [9] network without any bells and whistles as our baseline model. We then show qualitative and quantitative evidence that our feature binding procedure improves the network’s ability to segment highly occluded objects in complex scenes (Sec. 4.1.1), as well as objects found in out-of-context scenarios (Sec. 4.1.2). Throughout the experiments, we compare our method to recent mixing strategies, CutMix [38] and Mixup [39]. Although Mixup and CutMix did not explicitly design their strategies for dense labeling; however, in CutMix, the authors use CutMix and MixUp for image localization and object detection tasks, so we view their strategies as a general data augmentation technique. Next, we evaluate the robustness of our method to a variety of adversarial attacks (Sec. 4.2). Finally, we conduct an ablation study (Sec. 4.3) to better tease out the underlying mechanisms giving performance boosts by evaluating the various image blending strategies and network architectures.

Implementation Details. We implement our proposed feature binding method using Pytorch [6]. We apply bilinear interpolation to upsample the predictions before the losses are calculated. The *feature binding* network is trained using stochastic gradient descent for 30 epochs with momentum of 0.9, weight decay of 0.0005 and the “poly” learning rate policy [5] which starts at $2.5e^{-4}$. We use the same strategy during the denoising stage of training, but with an initial learning rate of $2.5e^{-5}$. During training, we apply random cropping to form 321×321 input images whereas testing is performed on the full resolution image.

4.1 Results on Semantic Segmentation

First, we show the improvements on segmentation accuracy by our method on the PASCAL VOC 2012 validation dataset. We present a comparison of different baselines and our proposed approach in Table 1. As shown in Table 1, feature binding improves the performance significantly more than other approaches [38, 39]. Following prior works [6, 30, 40], before evaluating our method on the test set, we first train on the augmented training set followed by fine-tuning on the original trainval set. As shown in Table 1, DeepLabv3 with feature binding achieves 82.1% mIoU which outperforms the baseline significantly. Sample predictions of our method and the baselines are shown in Fig. 3. As shown in Fig. 3, feature binding

# of Images	Occlusion		Number of Objects				Number of Unique Objects		
	1-Occ	All-Occ	1-Obj	2-Obj	3-Obj	4-Obj	2-Obj	3-Obj	4-Obj
# of Images	1128	538	695	318	167	98	375	121	23
DeepLabv3	75.5	74.9	74.6	74.8	76.0	72.1	72.5	63.5	64.8
DeepLabv3 + Mixup	75.4	72.3	77.9	74.3	71.7	69.6	72.0	58.1	55.4
DeepLabv3 + CutMix	76.4	74.3	78.3	75.4	73.0	72.1	72.3	60.1	56.3
DeepLabv3 + Binding	78.0	76.2	80.7	77.2	76.3	73.4	74.3	62.1	64.9

Table 2: Results on complex scenes in terms of mIoU, evaluated using various subsets from PASCAL VOC 2012 val set. *Occlusion*: Number of occluded objects in the image. *Number of Objects*: Number of objects in the image. *Number of Unique Objects*: Unique object classes contained in the image.

is very effective in capturing more distinct features for labeling occluded objects and plays a critical role in separating different semantic objects more accurately. Note the ability of our method to segment scenes with a high degree of occlusion, thin overlapping regions, or complex interaction between object categories. While other methods identify the dominant categories correctly, they often fail to relate the activations of smaller occluding features to the correct categorical assignments.

4.1.1 Segmenting Highly Occluded Objects in Complex Scenes

We argue that our mixing and source separation strategy is more powerful than other strategies in complex scenes with large amounts of occlusion. One reason for this is our mixing strategy (Sec. 3.1) blends images based on categorical clusters with dynamic blending ratios. This means that the network will see more images with a wide array of categories blended together, as every category is guaranteed to be blended with every other category. On the other hand, other strategies use two randomly selected images to blend. This means the statistics of the generated images will be largely driven by the statistics of the original dataset. Further, the SSM specifically is designed for separating features *before* the final layer of the network, allowing for finer details and semantics to be encoded into the target and phantom streams. For the other methods, they have a single prediction, which does not allow for these details to be separated early enough in the network to encode as much information as our method.

To substantiate this claim we evaluate each method under three specific data distributions that range in amount of occlusion and complexity: (i) *Occlusion*: at least one object has occlusion with any other objects (1-Occ) in an image and all objects have occlusion (All-Occ), (ii) *Number of Objects*: total number of object instances regardless of classes, and (iii) *Number of Unique Objects*: total number of unique semantic categories. The results are presented in Table 2. Our method outperforms the other mixing based methods in all cases. Note that the improvements on all occlusion and larger number of unique categories case are particularly pronounced for our binding model as the performance drop is significantly less than the other methods, when only considering images with many unique objects.

4.1.2 Segmenting Out-of-Context Objects

A model that heavily relies on context would not be able to correctly segment compared to the model that truly understands what the object is irrespective of its context. We argue that our mixing strategy performs better in out-of-context scenarios, as category-based mixing reduces bias in the dataset’s co-occurrence matrix. We conduct two experiments to quantitatively evaluate each method’s ability to segment out-of-context objects.

For the first experiment, we identify the top five categories that frequently co-occur with *person* based on the training set, since person has the most occurrences with all other cate-

	Co-occur with person					Exclusive				
	horse	mbike	bicycle	bottle	car	horse	mbike	bicycle	bottle	car
# of Images	32	34	30	20	45	44	23	29	35	45
DeepLabv3	87.9	81.6	77.7	89.7	89.7	90.9	91.5	60.4	85.4	96.0
DeepLabv3 + Mixup	86.9	82.8	76.5	87.6	86.2	92.5	93.0	60.0	80.6	95.5
DeepLabv3 + CutMix	86.2	83.6	76.0	87.4	87.9	94.1	93.8	61.3	82.6	96.2
DeepLabv3 + Binding	90.0	87.7	79.6	87.9	89.1	94.0	93.8	61.9	88.3	96.6

Table 3: mIoU results on the PASCAL VOC 2012 val set, for the co-occurrence of the most salient person category with five other categories and the results when these five categories appear alone.

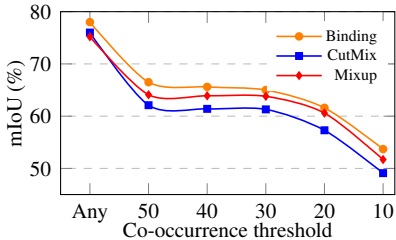


Figure 4: Performance on images with various levels of object co-occurrence. *Binding* performs better on subsets of images with unlikely co-occurrences.

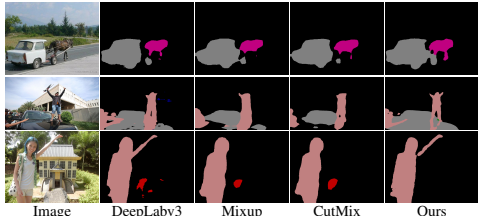


Figure 5: Qualitative examples on the (top row) Out-of-Context [16] and (bottom two rows) UnRel [22] datasets. Feature binding improves out-of-context performance.

gories based on the co-occurrence matrix. We report performance in Table 3 on two different subsets of data: (i) *Co-occur with Person*: images with both the person and object in it, and (ii) *Exclusive*: images with only the single object of interest. As can be seen from the table, when bottle co-occurs with person all the methods are capable of segmenting bottle and person precisely whereas the IoU for bottle is significantly reduced when bottle occurs alone. However, our proposed method successfully maintains performance on the exclusive case. For the second out-of-context experiment, we first create different subsets of images from the VOC 2012 val set based on the training set’s co-occurrence matrix. We select thresholds $\{50, 40, 30, 20, 10\}$, and only keep images which have objects that occur less than the chosen threshold. For instance, the threshold value 50 includes all the images where the co-occurrence value of object pairs is less than 50 (e.g., cat and bottle occur 18 times together, therefore images containing both will be in all subsets except the threshold of 10). Figure 4 illustrates the result of different baselines and our method with respect to co-occurrence threshold. Our method outperforms the baselines for all the threshold values.

We next perform a cross-dataset experiment by taking our model trained on the PASCAL VOC training set and evaluate on the publicly available Out-of-Context [16] and UnRel [22] datasets. Fig. 5 visualizes how the segmentation models trained with only VOC 2012 co-occurring objects performs when objects appear without the context seen in training. Even with such challenging images with out of context objects (person *on top* of car), our method produces robust segmentation masks while the baselines fails to segment the objects with detail. Since Out-of-Context and UnRel dataset do not provide segmentation ground-truth we cannot report quantitative results on these datasets.

4.2 Adversarial Robustness

We further claim our technique works as an implicit defense mechanism against adversarial images similar to [7, 8, 12, 17, 38, 39]. This is because the network optimization, in the

Networks	Clean	Adversarial Images				
		UAP [28]		GD-UAP [29]		
		ResNet	GNet	R-No	R-All	R-Part
DeepLabv3	75.9	59.1	63.6	58.7	56.8	56.5
+ Mixup	75.2	62.9	63.2	54.8	53.0	53.6
+ CutMix	76.2	60.9	64.3	47.4	46.4	46.9
+ Binding	78.0	68.7	70.2	63.9	63.1	62.8

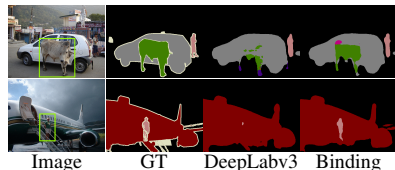


Table 4: The mean IoU for baselines and our approach when attacked with UAP [28] and GD-UAP [29]. **Figure 6:** Two challenging images where semantic objects are highly occluded.

form of source separation to solve the binding problem, enhances the capability of interacting with noisy features while imposing a high degree of resilience to interference from the superimposed image.

Adversarial Attacks. We generate adversarial examples using various techniques, including the Universal Adversarial Perturbation (UAP) [28] and Generalizable Data-free Universal Adversarial Perturbation (GD-UAP) [29] under different settings. We use publicly available perturbations of these methods to generate adversarial examples for the VOC 2012 val set. For UAP, which is a black-box attack, we generate adversarial images with both ResNet152 and GoogleNet based universal perturbations. GD-UAP is a grey-box attack, as it generates a perturbation based on the source data (VOC 2012 train set) and the backbone network (ResNet101). For GD-UAP, we compare different levels of adversarial attack strength by generating the perturbation based on various amounts of source data information.

Robustness of Segmentation Networks. We evaluate the robustness of different methods to adversarial examples and show how feature binding-driven training learns to significantly mitigate performance loss due to perturbation. Table 4 shows the robustness of different baselines and our approach on the PASCAL VOC 2012 validation dataset. In general, DeepLab-based methods [5] achieve higher mIoU for the segmentation task on clean examples and is also shown to be more robust to adversarial samples compared to the shallower networks [10]. In the case of black-box attacks, the adversarial examples originally generated by UAP on ResNet152, are less malignant (68.7% mIoU) when the feature binding concept is applied, while being effective in significantly reducing the performance of other methods.

When we apply a semi-white-box attack under the setting (R-All), where VOC 2012 training data and the ResNet101 network are used to generate the perturbation, DeepLabv3 and Mixup show robustness against adversarial examples which is improved by applying feature binding. Surprisingly, the performance of CutMix is significantly reduced when tested against adversarial samples generated by GD-UAP. Similarly, we find that DeepLabv3, Mixup, and CutMix are also vulnerable to adversarial cases under the *R-No* and *R-Part* settings, where no data and partial data is used respectively to generate the perturbations. Notably, DeepLabv3+Binding exhibits significant robustness to extreme cases which further reveals the importance of feature binding to successfully relate internal activations corresponding to common sources in the adversarial images.

4.3 Ablation Studies

In this section, we examine the possible variants of our feature binding pipeline by considering three different settings. Note that for all the experiments, except the denoising in our ablation study, we choose ResNet101 based distributed gating network [24] as the backbone.

Method	Image Blending Techniques						Methods		Methods		mIoU	
	No	\mathcal{C}_{f_b}	\mathcal{R}_{f_b}	$\mathcal{C}a_{f_b}$	$W\mathcal{R}_{f_b}$	\mathcal{M}_{f_b}	DeepLabv3	75.9	DIGNet [24]	75.1	+ Ours (w/o FBH)	75.3
DIGNet [24]	75.1	76.1	74.5	74.5	69.7	75.4	+ ours (w/ DN)	76.2	+ Ours (w/ FBH)	76.1	+ Ours (w/ FBH)	76.1

(a)

(b)

(c)

Table 5: (a) Performance comparison of different blending techniques on the VOC 2012 val set. \mathcal{C}_{f_b} : Clustering based blending discussed in Sec. 3.1, \mathcal{R}_{f_b} : Each sample randomly paired with 10 samples, $\mathcal{C}a_{f_b}$: Within category random pair, $W\mathcal{R}_{f_b}$: random pairing with fixed $\delta = 0.6$, \mathcal{M}_{f_b} : Random pairs from half of the train set and using standard images from the other half. (b) Significance of feature denoising stage (DN). (c) Performance comparison with and without the feature binding head (FBH).

Feature Binding Driven Blending Techniques. We report the labeling results of several blending techniques in Table 5(a). If we select two images randomly and allow the two images to be any class (\mathcal{R}_{f_b}), the performance is lower than the proposed clustering based technique. Additionally, the performance was not improved when we mix two images belonging to the same category ($\mathcal{C}a_{f_b}$). However, our proposed clustering based blending, \mathcal{C}_{f_b} , achieves higher mIoU compared to possible alternatives highlighting the importance of the choice of pattern collisions in applying feature binding.

Feature Denoising and Feature Binding Head. We examine the effectiveness of feature denoising (DN) stage and report results in Table 5(b). We also conduct experiments (Table 5(c)) varying the source separator module, including the feature binding head (FBH). The overall performance can be improved with the addition of a feature denoising stage and feature binding head, see Table 5 (b) and (c), respectively. We believe the feature binding head allows the network to make a more informed final prediction based on the source *and* the phantom activations, and therefore learns to identify harmful features at inference time, leading to a more accurate prediction.

5 Discussion and Conclusion

Training with the feature binding pipeline enables learning resilient features, separating sources of activation, and resolving ambiguity with richer contextual information. Although DeepLabv3 is a powerful segmentation network, there are cases (see Fig. 6) where background objects are correctly classified (car and plane) but other semantic categories are not separated correctly due to high degrees of occlusion (person on the stairs, see Fig. 6 right). In contrast, the feature binding based learning approach is highly capable of resolving such cases by learning to separate source objects and tying them to specific regions.

In summary, we have presented an approach to training CNNs based on the notion of feature binding. This process includes, as one major component, careful creation of categorical collisions in data during training. This results in improved segmentation performance, and also promotes significant robustness to adversarial perturbations. Denoising in the form of fine-tuning shows further improvement along both these dimensions.

Acknowledgements: The authors gratefully acknowledge financial support from the Canadian NSERC Discovery Grants, Ontario Graduate Scholarship, and Vector Institute Postgraduate Affiliation award. K.G.D. contributed to this work in his personal capacity as an Associate Professor at Ryerson University. We also thank the NVIDIA Corporation for providing GPUs through their academic program.

References

- [1] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018.
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *TPAMI*, 2017.
- [3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.
- [4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.
- [5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018.
- [6] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 2012.
- [7] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. *arXiv preprint arXiv:2007.03943*, 2020.
- [8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.
- [9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
- [10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.
- [11] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916*, 2019.
- [12] Pando Georgiev, Fabian Theis, and Andrzej Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *TNN*, 16(4):992–996, 2005.
- [13] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *ECCV*, 2016.
- [14] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, and Adam Prügel-Bennett Jonathon Hare. Fmix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020.
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

-
- [16] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *ASLP*, 2015.
- [17] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv:1801.02929*, 2018.
- [18] Md Amirul Islam, Shujon Naha, Mrigank Rochan, Neil Bruce, and Yang Wang. Label refinement network for coarse-to-fine semantic segmentation. *arXiv:1703.00551*, 2017.
- [19] Md Amirul Islam, Mrigank Rochan, Neil D. B. Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *CVPR*, 2017.
- [20] Md Amirul Islam, Mahmoud Kalash, and Neil D.B. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*, 2018.
- [21] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Semantics meet saliency: Exploring domain affinity and models for dual-task prediction. In *BMVC*, 2018.
- [22] Md Amirul Islam, Mrigank Rochan, Shujon Naha, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for coarse-to-fine dense semantic image labeling. *arXiv preprint arXiv:1806.11266*, 2018.
- [23] Rezaul Karim, Md Amirul Islam, and Neil DB Bruce. Recurrent iterative gating networks for semantic segmentation. In *WACV*, 2019.
- [24] Rezaul Karim, Md Amirul Islam, and Neil DB Bruce. Distributed iterative gating networks for semantic segmentation. In *WACV*, 2020.
- [25] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative instance segmentation. In *CVPR*, 2016.
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.
- [29] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *TPAMI*, 2018.
- [30] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.
- [31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

- [32] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *ICCV*, 2017.
- [33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.
- [34] Stewart Shipp, Daniel L Adams, Konstantinos Moutoussis, and Semir Zeki. Feature binding in the feedback layers of area v2. *Cerebral cortex*, 19(10):2230–2239, 2009.
- [35] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don’t judge an object by its context: Learning to overcome contextual bias. *arXiv preprint arXiv:2001.03152*, 2020.
- [36] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *CVPR*, 2018.
- [37] Anne Treisman. Feature binding, attention and object perception. *Phil. Trans. R. Soc. B.*, 1998.
- [38] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.
- [39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
- [40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.